Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

A HEAD-AND-TORSO MODEL FOR LOW-FREQUENCY BINAURAL ELEVATION EFFECTS

Carlos Avendano and V. Ralph Algazi

CIPIC
U.C. Davis
Davis, CA 95616, USA
{carlos,algazi}@ece.ucdavis.edu

Richard O. Duda

Department of Electrical Engineering
San Jose State University
San Jose, CA 95192, USA
rod@duda.org

ABSTRACT

Low-frequency elevation-dependent features appear in HRTF measurements because of torso and shoulder reflections and head diffraction effects. A simple structural model that accounts for these features is presented. Listening tests show that the model produces significant elevation cues for virtual sound sources whose spectra are limited to frequencies below 3 kHz. The low-frequency binaural elevation cues are perceptually significant away form the median plane, and complement high-frequency monaural pinna cues.

1. INTRODUCTION

It is well known that the static localization of a sound source in azimuth and elevation involves both binaural and monaural cues [1]. As Lord Rayleigh demonstrated, localization in azimuth is dominated by two binaural cues — the interaural time difference (ITD) due to propagation delay across the head, and the interaural level difference (ILD) due to head diffraction. A simple spherical head model can explain the phenomena and can be used to produce quite convincing azimuth effects [2].

Localization in elevation is generally thought to be determined by monaural spectral cues due to pinna diffraction [1]. Because the effects of diffraction appear only when the wavelength becomes comparable to or smaller than the diffracting object, pinna diffraction is frequency dependent, and is not significant for frequencies below 3 or 4 kHz [3, 4]. Thus, it is generally believed that localization in elevation requires a wide-bandwidth source with substantial high-frequency energy.

However, we have discovered that there are also important low-frequency binaural elevation cues due to multipath head-diffraction effects and shoulder and torso reflections, and that these cues are surprisingly effective with low-frequency sources.

This discovery occurred in the course of work on developing structural models for head-related transfer functions (HRTFs) [5, 6]. The physical sources of these cues are most easily seen in the head-related impulse responses (HRIRs), where they appear as localized events that are easier to understand than the corresponding pattern of spectral ripples in HRTF magnitude data.

After we had created a simple structural model that seemed to account for the major properties of these physical features, we tested it informally by listening to monaural sound signals filtered by the model. Using test signals with no significant energy content above 3 kHz, we found that the model had a strong effect on apparent source location away from the median plane but little effect near the median plane. This convinced us that there are important low-frequency binaural cues for elevation, and encouraged us to test the model more systematically.

To present these results, we begin by examining image representations of HRTFs and HRIRs that reveal the head, shoulder and torso contributions. We next describe our model, and compare its response to the measured data. We then describe our listening tests, and conclude with a discussion on the significance of these results for synthesizing spatial sound.

2. MULTIPATH COMPONENTS OF THE HRTF

In measuring HRTFs, we use an interaural-polar coordinate system in which the elevation angle ϕ is the angle between the horizontal plane and a plane through both the interaural axis and the source. With this coordinate system, a surface of constant azimuth θ is what Woodworth calls a "cone of confusion."

Fig. 1 shows the elevation dependence of the HRTF for the KEMAR manikin around a $\theta = 45°$ cone of confusion. The response magnitude (in dB) corresponds to brightness. The relatively greater brightness for the ipsilateral ear at high frequencies reveals the effects of head shadow. Features that change with elevation are potential elevation cues. These include the so-called "pinna notches" that appear as dark streaks between about 6 and 12 kHz.

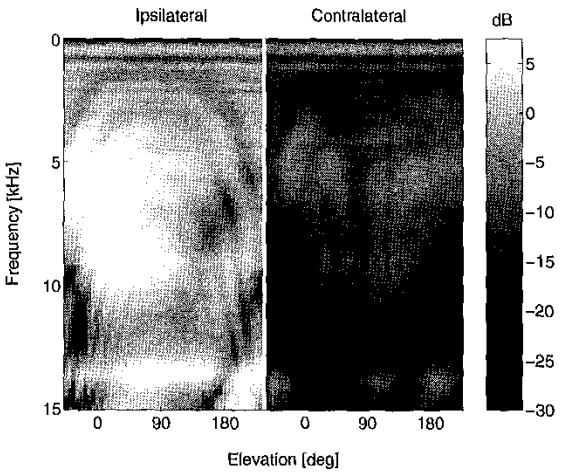

Figure 1: *Magnitude of the HRTF of KEMAR at* $\theta = -45°$.

However, one can also see arch-shaped notches at frequencies as low as 1 kHz. These features give rise to the elevation-dependent low-frequency ILD shown in Fig. 2, which becomes basically uniform when the torso is removed. The pinnae have essentially no effect on these features, as can be seen from the comparison of the KEMAR ILD with and without pinnae (see Fig. 2).

The various components responsible for this behavior at low frequencies can be more easily seen in the time domain. An image representation of the HRIRs corresponding to the KEMAR data with no pinnae is shown in Fig. 3. Here each column corresponds to an impulse response at a particular elevation angle. The initial wavefront appears as the white band near the top of each image (to reveal the detail in the impulse response, large values were limited, which is the reason for the uniform brightness of the initial pulse). The initial reflections from the chest and back appear as the two V-shaped contours that are most clearly seen in image for the ipsilateral ear. None of these features is present in the HRIR when the torso is removed. The maximum time delay occurs when the source is overhead, and is approximately 0.8 ms. This corresponds to a distance of about 27 cm, which is roughly twice the distance from the ear canal to the shoulders.

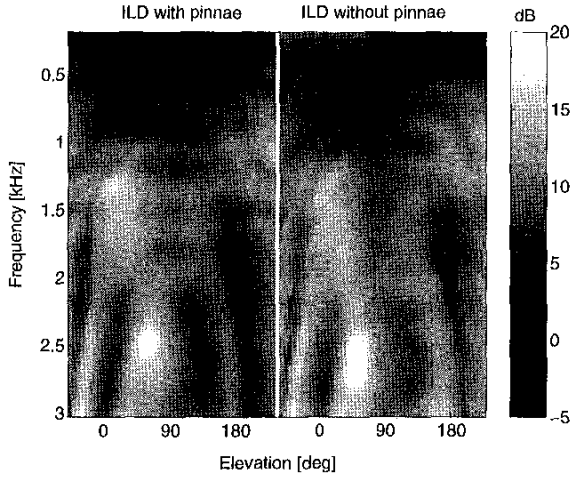

Figure 2: *ILD of KEMAR with pinnae and without pinnae.*

The impulse response for the contralateral ear is considerably more complex than that for the ipsilateral ear. In addition to the two V-shaped contours for chest and back reflections, we observe two other prominent contours that merge into the initial pulse at vertical elevations (see Fig. 3). These correspond to the two primary paths from the source to the contralateral ear, one around the front of the head and one around the back of the head [8]. Finally, we note that the arrival time for the initial pulse varies with elevation, and is greatest when the source is overhead. Thus, the interaural time delay (ITD) is not actually constant around a cone of constant azimuth, but varies with elevation. This behavior can be explained by the fact that (a) the head is more nearly ellipsoidal than spherical, and (b) the ears are displaced downward and toward the back [7]. At high frequencies, the detailed structure of the contralateral response are obscured by head shadow, and does not seem to be important [6]. However, the presence of the additional head diffraction components has a significant influence on the low-frequency ILD, and cannot be ignored.

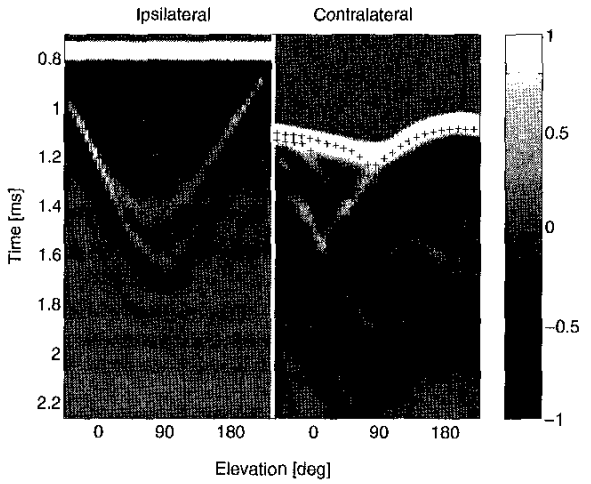

Figure 3: *HRIR of KEMAR without pinnae at $\theta = -45°$. The "+" symbols show the delay data points used in the model.*

3. A STRUCTURAL MODEL OF THE HEAD-AND-TORSO

A complete structural model of the HRTF should contain components to account for all of the multiple paths by which sound waves can reach the ears. In this paper, we present a greatly simplified model that ignores pinna resonances, and models the head-and-torso impulse response by a combination of scale factors k, delays T, and simple filters H.

The model for the ipsilateral ear is shown in the left half of Fig. 4. The signal X from the source produces the ipsilateral output X_i. X_{Hi} represents the wave that is incident on the ipsilateral ear. The head filter H_{Hi} is a single-pole, single-zero spherical-head approximation that boosts the high-frequency response on the ipsilateral side [5].

The incident wave is the sum of three components, one due to the direct path from the source, one due to a reflection from the front of the torso, and one due to a reflection from the back of the torso. The strengths and time delays associated with the reflected components vary with elevation, and are represented by the four parameters k_{fi}, k_{bi}, T_{fi} and T_{bi} (see Fig. 4). The scale factor k_{fi} for the front reflection is large for frontal elevations, and drops off as the source moves around to the back. Similarly k_{bi} (which accounts for the shoulder reflection as well as the back reflection) increases as the source moves around to the back. The two time delays are measured from the initial arrival time, and the elevation dependence of all four parameters was determined by fitting spline curves to the data shown in Fig. 3. Gains were computed based on the power of the HRIR in a short window (0.2 ms) around the reflection.

A closer examination of the torso components of the HRIR reveals that they have a narrower bandwidth than the initial pulse. This is because torso "reflections" are actually fairly complex diffraction phenomena, and diffraction is frequency dependent. The torso filter H_T accounts for the broadening of the torso reflection pulses. Based on our experimental data, H_T was modeled as a simple single-pole single-zero low-pass filter with a 675-Hz cutoff frequency.

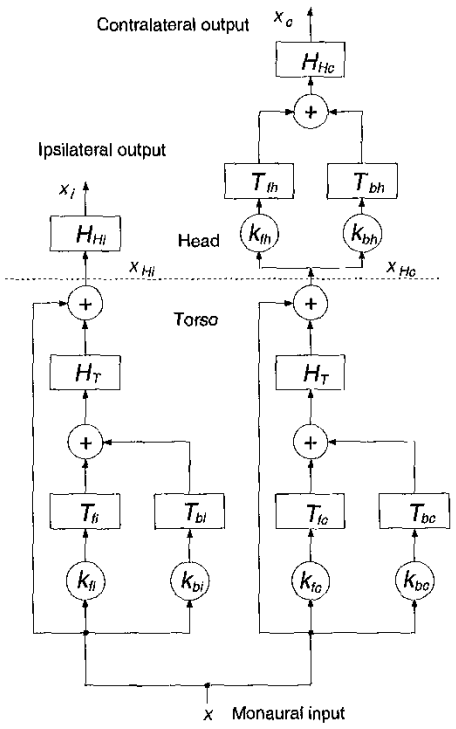

Figure 4: *A structural HRTF model. The monaural input reaches the head by a direct path and via delayed torso reflections, which depend on both azimuth and elevation. The head shadow and delay depend on azimuth only.*

The right half of Fig. 4 shows the model for the contralateral ear. It is basically the same as the model for the ipsilateral ear, except for the addition of the time delays T_{fh} and T_{bh} and the scale factors k_{fh} and k_{bh} for the two paths around the front and the back of the head. All four of these head parameters vary with elevation, and their values were determined by fitting spline curves to the data shown in Fig. 3.

4. LISTENING TESTS

Our initial observations were that the model provided a good sense of elevation for low frequency sources ($f_c < 3$ kHz) away from the median plane. To evaluate the model more systematically, we performed absolute localization listening tests where the task was simply to report the perceived elevation angle. We also tested a modified model in which the torso delays and relative gains were the same for both ears, so that there were monaural spectral cues but there was no elevation variation in the ILD.

Both models were tested at 50 different elevations ($\phi = -45°$ to $\phi = 235°$ in $5.625°$ increments) around the cone of confusion at $\theta = -45°$. The test was divided into two sessions so that locations to the front ($\phi < 90°$) were tested separately from locations to the back ($\phi \geq 90°$), and subjects were notified of the hemisphere for which they were being tested. This was done to decrease the influence of front/back confusions in the subjects' judgment.

In each session the subject was presented with 200 test trials where location and models (either with ILD and without ILD) were randomly selected. Thus, each subject performed 400 evaluations where, on the average, each location was tested two times for each model. A total of four subjects participated in the tests.

The stimulus used in the tests was a 40 -Hz amplitude modulated white noise burst with a duration of 500 ms. Before presentation, this stimulus was band-limited to 3 kHz by a 10^{th} order Butterworth filter and repeated four times with a gap of 500 ms between bursts. A single set of headphones (AKG K240-DF) was used for all tests.

A graphical user interface was used to collect the subjects' responses. The cone of confusion under test, projected on the median plane, was shown schematically as a circle with a number of marked elevation sectors. Front, back, up and down directions were labeled to define orientations on that circular diagram. After listening to the stimulus (one presentation) the subjects were asked to select on the diagram the perceived angle. The results for one of the subjects (S3) are illustrated in Fig. 5. Both the correlation coefficient r and slope s of the best fitting line are shown. Precise and accurate localization would result in both coefficients equal to unity. The results of the tests for all subjects are summarized in Table 1 and Table 2.

Table 1: Slope s

Subject	Front ILD	Front no ILD	Back ILD	Back no ILD
S1	0.79	0.33	1.04	0.38
S2	0.42	0.30	0.69	0.55
S3	0.90	0.47	0.51	0.34
S4	0.69	0.46	0.34	0.31

Table 2: Correlation Coefficient r

Subject	Front ILD	Front no ILD	Back ILD	Back no ILD
S1	0.65	0.36	0.77	0.45
S2	0.35	0.26	0.68	0.54
S3	0.77	0.45	0.62	0.56
S4	0.44	0.29	0.25	0.25

The low-frequency head-and-torso cues that we have captured in our model contain only part of the elevation information that is in the true HRTF. We note however that the binaural model may provide quite strong elevation information with a high slope and high correlation. Monaural cues are also present, but are substantially weaker and more uncertain. We also observe that there is a substantial variation of the results from subject to subject. Since model parameters were extracted from the KEMAR experimental data and used for all subjects, it is possible that improved performance would result from customizing model parameters to subjects.

5. DISCUSSION

In an earlier work, we had noted that the ILD varies with elevation as well as azimuth, and had speculated that these differences might provide important elevation cues [9]. However, that study focused on high-frequency features due to pinna diffraction, and overlooked the low-frequency ILD. Similarly, in a previous model

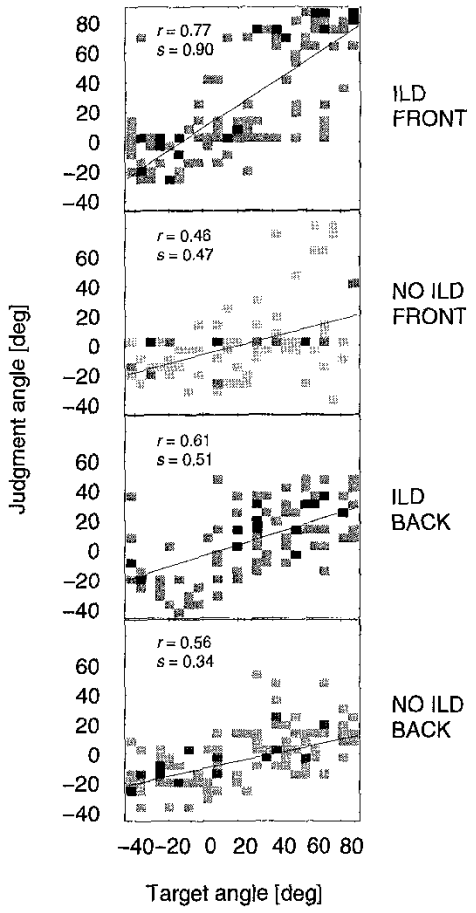

Figure 5: *Perceptual evaluation results for subject S3. The number of responses in a particular location is represented by the gray level (lighter squares represent a single response).*

we allowed for shoulder reflections, but we assumed that they would have little effect on elevation perception and did not include shoulder reflections in the formal evaluation [5]. Although a more thorough evaluation of the current model is needed, we now have no doubt that torso effects provide significant elevation cues, and that they do not require that the source have energy above 2 or 3 kHz.

These results have several implications. First, they explain why many binaural recordings that are made with a dummy head but no torso produce elevated images. Introducing the proper torso cues should help bring the sound images down to their proper elevation.

Second, they raise a warning about measured HRTFs. To model a point source many HRTF measurements (including our own) are made using small loudspeakers. Because it is difficult to compensate for their low output at low frequencies, it is tempting to extrapolate to the "known" theoretical solution. However, this must be done carefully to prevent losing the low-frequency torso cues.

Third, they suggest that it is not necessary to have wide-band signals and excellent high-frequency hearing to judge the elevation of a source. Even telephone-bandwidth signals have the potential to provide useful cues.

There are a number of unanswered questions. Do the significant person-to-person variation in the torso cue have a strong effect on elevation perception? Other unanswered questions concern the effects of body movements, such as turning the head relative to the torso, bending over, and moving from seated to standing. Accounting for these dynamic changes could be important to localization with low frequencies. Finally, it is also unclear whether or not these effects can be effectively produced using crosstalk-canceled stereo, which is difficult to do well at low frequencies.

We conclude by observing that the head-and-torso reflections complement pinna cues, and that a full HRTF model must contain pinna models. The combination of all consistent cues is surely important for the synthesis of convincing spatial sound. However, just as one cannot ignore the pinna cues, we now know that one should not ignore the torso cues either.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant NSF IRI-96-19339. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the National Science Foundation. The authors would like to thank Dennis Thompson for his help in collecting the data, and the subjects who participated in the listening tests.

7. REFERENCES

[1] Blauert, J., *Spatial Hearing*, MIT Press, Cambridge, MA, 1983.

[2] Kuhn, G. F., "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, 62, pp. 157-167, 1977.

[3] Shaw, E. A. G., "Acoustical features of the human external ear," in *Binaural and Spatial Hearing in Real and Virtual Environments*(R. H. Gilkey and T. R. Anderson, Eds.). Mahwah, NJ: Lawrence Erlbaum Associates, 1997. pp. 49-75.

[4] Kuhn, G. F.,"Acoustics and measurements pertaining to directional hearing," in *Directional Hearing*(W. A. Yost and G. Gourevitch, Eds.). New York, NY: Springer Verlag, 1987. pp. 3-25.

[5] Brown, C. P. and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech and Audio Proc.* 6, pp. 476-488, 1998.

[6] Avendano, C., R. O. Duda and V. R. Algazi, "Modeling the contralateral HRTF," presented at the Audio Engineering Society 16th International Conference on Spatial Sound Reproduction (Rovaniemi, Finland, April 1999).

[7] Duda, R. O., C. Avendano and V. R. Algazi,"An adaptable ellipsoidal head model for the interaural time difference," *Proc. ICASSP99*, 2, Paper 1592 (Phoenix, AZ, March 1999).

[8] Duda, R. O. and W. L. Martens, W. L., "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, 104, pp. 3048-3058, 1998.

[9] Duda, R. O., "Elevation dependence of the interaural transfer function," in *Binaural and Spatial Hearing in Real and Virtual Environments*(R. H. Gilkey and T. R. Anderson, Eds.). Mahwah, NJ: Lawrence Erlbaum Associates, 1997. pp. 49-75.